Response to Reviewers " Nonexistence of unbiased test for high-dimensional linear model and a Bayesian-motivated test "

Rui Wang[1] and Xingzhong Xu[*1,2]

[1] School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081,China
[2] Beijing Key Laboratory on MCAACI, Beijing Institute of Technology, Beijing 100081,China

Wednesday 2nd October, 2019

We thank both reviewer for their helpful comments and critiques. We have carefully considered the comments and made corresponding changes to the paper. Our responds are as follows.

1   Response to reviewer 1

1. The manuscript should provide comparison with the literature that exploits sparsity because under the null hypothesis the model parameter is sparse (with at most q non-zero entries). For example, Zhang and Cheng (2017) studies the same problem of testing (high-dimensional components) of linear regressions. Apart from assuming random designs, this literature attempts to detect deviations in l_∞-norm, while the current manuscript seems to aim at detecting deviations in l_2-norm. In my opinion, more discussion on this literature would be helpful so that the reader would get a fair impression on what has been done and what assumptions different methods require. For example, here symmetry of error distribution is needed (Assumption 1), which is not usually required in the literature on high-dimensional inference.

Answer: Following the reviewer's suggestion, we have provided the simulation results of the desparsifying lasso test of Zhang and Cheng (2017) (abbreviated as DL) and add the following comment to clarify the application regime of the proposed test.

> In the sparse $\boldsymbol{\beta}_b$ setting of Model II and Model III, the proposed test is not as powerful as the DL test. In fact, as an extreme value type test, the DL test is designed for

---
[*]Corresponding author
Email address: xuxz@bit.edu.cn

sparse linear models. In contrast, the proposed test test is a quadratic form which offers higher power when the signal is dense. On the basis of the simulations, we can recommend using the proposed test when there is no prior knowledge of sparsity.

As the reviewer pointed out, the symmetry of error distribution is not usually required in the literature on high-dimensional inference. However, this condition is a key assumption in the asymptotic theory of quadratic forms and is also assumed by Bai et al. (2018). It will be a highly nontrivial task to relax this assumption. We leave it for possible future work. We add the following comment in Section 5:

> Our Theorem 2 gives a general approximation for the distributions of quadratic forms. A key assumption in Theorem 2 is that the distribution of ξ_1 is symmetric about 0.. This condition is also assumed by Bai et al. (2018) in the study of central limit theorem of quadratic form. However, this condition is not usually required in the literature on high-dimensional inference. We think it will be an interesting and useful work to relax the symmetric condition in Theorem 2.

2. Theorem 4 provides the power analysis, but its assumption in Equation (8) needs more discussion. For example, at the end of Section 3, the manuscript uses the factor model to illustrate the power of the proposed test over Goeman et al. (2006). However, this discussion is based on Theorem 4 and my rough calculation suggests that Equation (8), which is required by Theorem 4, does not seem to hold for factor models. Let me outline some details.

(a) **Supppose that $q = 0$ (so $\tilde{\mathbf{U}}_a = \mathbf{I}_n$) and $\mathbf{X}_b = \begin{pmatrix} W_1^\top \\ \vdots \\ W_n^\top \end{pmatrix}$ with $W_i = BF_i + u_i$, where F_i is a scalar from i.i.d. $\mathcal{N}(0,1)$, u_i is from $\mathcal{N}(0, \mathbf{I}_p)$ and B is from $\mathcal{N}(0, \mathbf{I}_p)$. Assume F_i, u_i and B are independent.**

(b) **Since $\mathrm{E}(\gamma_I^{2k}) \geq \mathrm{Var}(\gamma_I^k)$, a necessary condition for Equation (8) is**

$$\frac{(\gamma_1^k - \mathrm{E}_I(\gamma_I^k))^2}{n\,\mathrm{E}_I(\gamma_I^{2k})} \to 0,$$

where E_I denotes expectation taken over randomness of I.

(c) **Based on the computation above Theorem 4 (i.e. $\mathrm{E}_I(\gamma_I^k) = \mathrm{tr}(\mathbf{X}_b \mathbf{X}_b^\top)^k / n$), we know that $\mathrm{E}_I(\gamma_I) = n^{-1} \sum_{i=1}^n W_i^\top W_i$ and**

$$\mathrm{E}_I(\gamma_I^2) = n^{-1} \sum_{i=1}^n (W_i^\top W_i)^2 + n^{-1} \sum_{i=1}^n \sum_{j=1, j \neq i}^n (W_i^\top W_j)^2.$$

(d) **For $k = 1$, we use the assumptions in (a) to determined the size of γ_1, $\mathrm{E}_I(\gamma_I)$ and $\mathrm{E}_I(\gamma_I^2)$. Simple computations would give use $\mathrm{E}\, W_i^\top W_i = 2p$, $\mathrm{E}(W_i^\top W_i)^2 = 6p^2 + 10p$ and**

2

$\mathrm{E}(W_i^\top W_j)^2 = p^2 + 5p$ **for** $i \neq j$. **Thus,** $\gamma_1 = O_p(np)$, $\mathrm{E}_I(\gamma_I) = O_p(p)$ **and** $\mathrm{E}_I(\gamma_I^2) = O_p(np^2)$.
Thus, $\frac{(\gamma_I^k - \mathrm{E}_i(\gamma_I^k))^2}{n\,\mathrm{E}_I(\gamma_I^{2k})}$ **with** $k = 1$ **seem to be of the order** $O_P(1)$ **instead of** $o_P(1)$.

(e) Overall, I think discussions on why Equation (8) should hold is needed, at least for the examples used at the end of Section 4 (i.i.d. rows and factor models).

Answer: We follow the reviewer's suggestion and add discussions on the condition (8) (which is Equation (9) in the revised paper). This condition corresponds to the condition $\lambda_1(\mathbf{A}^2)/\mathrm{tr}(\mathbf{A}^2) \to 0$ in Proposition 2, which is essential for the asymptotic normality of the statistics. We require the asymptotic normality so that we can give tractable asymptotic power functions. We add the following remark for Proposition 2:

> In Proposition 2, the condition $\lambda_1(\mathbf{A}^2)/\mathrm{tr}(\mathbf{A}^2) \to 0$ is equivalent to (6). To see this, note that under the normal distribution, the condition (6) becomes $\mathrm{tr}(\mathbf{A}^4)/\mathrm{tr}^2(\mathbf{A}^2) \to 0$. Then the equivalence of the two conditions follows from equality
>
> $$\frac{\lambda_1^2(\mathbf{A}^2)}{\mathrm{tr}^2(\mathbf{A}^2)} = \frac{\lambda_1(\mathbf{A}^4)}{\mathrm{tr}^2(\mathbf{A}^2)} \leq \frac{\mathrm{tr}(\mathbf{A}^4)}{\mathrm{tr}^2(\mathbf{A}^2)} \leq \frac{\lambda_1(\mathbf{A}^2)\,\mathrm{tr}(\mathbf{A}^2)}{\mathrm{tr}^2(\mathbf{A}^2)} = \frac{\lambda_1(\mathbf{A}^2)}{\mathrm{tr}(\mathbf{A}^2)}.$$
>
> As noted in Remark 2, the condition (6) is almost necessary for the asymptotic normality of quadratic forms. Hence this condition allows us to derive simple forms of asymptotic power functions. If this condition is violated, the distributions of noncentral quadratic forms are rather complicated and the asymptotic power functions will be intractable.

The reviewer found an error of our original paper through his/her careful analysis and elegant derivation. Although there is a minor typo in the reviewer's derivation ($(\mathrm{E}(W_i^\top W_i)^2$ should equal to $6p^2 + 12p$ rather than $6p^2 + 10p$), his/her conclusion is valid, that is, our Theorem 4 can not be applied to the factor model. Hence it is an error to discuss factor model based on the conclusion of Theorem 4. In the original paper, we did not realize this error and we are sorry about that.

We tried to remedy this error by considering a factor model with a divergent number of latent factors, as detailed below.

(a) Supppose that $q = 0$ (so $\tilde{\mathbf{U}}_a = \mathbf{I}_n$) and $\mathbf{X}_b = \begin{pmatrix} W_1^\top \\ \vdots \\ W_n^\top \end{pmatrix}$ with $W_i = BF_i + u_i$, where F_i is a m dimensional random vector with i.i.d. $\mathcal{N}(0,1)$ entries, u_i is from $\mathcal{N}(0,\mathbf{I}_p)$ and B is a $p \times m$ random matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Assume F_i, u_i and B are independent.

(b) As the reviwer pointed out, if m is fixed, the condition

$$\frac{\max_{1 \leq i \leq n}(\gamma_i - \mathrm{E}(\gamma_I))^2}{n\,\mathrm{Var}(\gamma_I)} \to 0$$

does not hold. We would like to show that if $m \to \infty$ and $m/n \to 0$, then the above condition will hold.

3

(c) We have

$$\mathrm{E}_I(\gamma_I) = n^{-1} \sum_{i=1}^{n} W_i^\top W_i,$$

and

$$\mathrm{Var}_I(\gamma_I)$$
$$= (n^{-1} - n^{-2}) \sum_{i=1}^{n} (W_i^\top W_i)^2 + n^{-1} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} (W_i^\top W_j)^2 - n^{-2} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} (W_i^\top W_i)(W_j^\top W_j).$$

(d) We can compute the following expectations:

$$\mathrm{E}\, W_i^\top W_i = \mathrm{E}(F_i^\top B^\top B F_i + 2 u_i^\top B F_i + u_i^\top u_i) = mp + 0 + p = (m+1)p.$$

$$
\begin{aligned}
\mathrm{E}(W_i^\top W_i)^2 &= \mathrm{E}(F_i^\top B^\top B F_i + 2 u_i^\top B F_i + u_i^\top u_i)^2 \\
&= \mathrm{E}(F_i^\top B^\top B F_i)^2 + 4\,\mathrm{E}(u_i^\top B F_i)^2 + \mathrm{E}(u_i^\top u_i)^2 + 2\,\mathrm{E}(F_i^\top B^\top B F_i u_i^\top u_i) \\
&= \mathrm{E}[\mathrm{E}[(F_i^\top B^\top B F_i)^2 | B]] + 4mp + (p^2 + 2p) + 2\,\mathrm{E}(F_i^\top B^\top B F_i)\,\mathrm{E}(u_i^\top u_i) \\
&= \mathrm{E}[2\,\mathrm{tr}(B^\top B)^2 + (\mathrm{tr}(B^\top B))^2] + 4mp + (p^2 + 2p) + 2mp^2 \\
&= \left\{ 2\left[mp^2 + m(m+1)p \right] + (m^2 p^2 + 2mp) \right\} + 4mp + (p^2 + 2p) + 2mp^2 \\
&= (m^2 + 4m + 1)p^2 + (2m^2 + 8m + 2)p.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(W_i^\top W_j)^2 &= \mathrm{E}(F_i^\top B^\top B F_j + u_i^\top B F_j + u_j^\top B F_i + u_i^\top u_j)^2 \\
&= \mathrm{E}(F_i^\top B^\top B F_j)^2 + \mathrm{E}(u_i^\top B F_j)^2 + \mathrm{E}(u_j^\top B F_i)^2 + \mathrm{E}(u_i^\top u_j)^2 \\
&= \left[mp^2 + m(m+1)p \right] + mp + mp + p \\
&= mp^2 + (m^2 + 3m + 1)p.
\end{aligned}
$$

Then we have $\mathrm{E}\left[\mathrm{E}_I(\gamma_I)\right] = (m+1)p$ and $\mathrm{Var}\left[\mathrm{E}_I(\gamma_I)\right] = O(mp^2)$. Hence $\mathrm{E}_I(\gamma_I) = mp(1+o_P(1))$. Also, we have

$$\mathrm{Var}_I(\gamma_I) = O_P(m^2 p^2) + n^{-1} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} (W_i^\top W_j)^2 + O_P(m^2 p^2)$$

It can be seen that $\mathrm{E}[n^{-1} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} (W_i^\top W_j)^2] = mnp^2(1 + o(1))$. With much more effort (by computing higher moments; see, e.g., Chen et al. (2010), Proposition A.2.(ii)), it may be shown that $\mathrm{Var}[n^{-1} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} (W_i^\top W_j)^2] = o(m^2 n^2 p^4)$. Thus, $\mathrm{Var}_I(\gamma_I^2) = mnp^2(1+o_p(1))$. Using some external knowledge of factor model or spiked covariance model (see, e.g., Wang and Xu (2020), Proposition 1), it can be shown that $\gamma_1 = O_P(np)$. In this case,

$$\frac{\max_{1 \leq i \leq n}(\gamma_i - \mathrm{E}(\gamma_I))^2}{n\,\mathrm{Var}(\gamma_I)} \leq \frac{2\gamma_1^2 + 2(\mathrm{E}(\gamma_I))^2}{n\,\mathrm{Var}(\gamma_I)} = O_P(1/m) \to 0.$$

4

We have shown that the condition

$$\frac{\max_{1\leq i\leq n}(\gamma_i^k - \mathrm{E}(\gamma_I^k))^2}{n\,\mathrm{Var}(\gamma_I^k)} \to 0$$

holds for $k = 1$. However, some external knowledge is used. Worth still, the above derivation can not be applied for $k = -1$ since it is hard to directly computing the moments of $\mathrm{E}(\gamma_I^{-1})$ and $\mathrm{Var}(\gamma_I^{-1})$. To verify the condition holds for $k = -1$, more advanced random matrix theory may be necessary, which deviates too much from the topic of our paper. Hence in the revised paper, we remove the discussion about the factor model. Instead, we add more discussion on the local power of the proposed test and the test of Goeman et al. (2006), as detailed in the answer of the next question.

3. I think the power comparison with Goeman et al. (2006) should be more formal. Currently, there is a comparison on when $\mathrm{Cov}(-\gamma_I^{-1}, \gamma_I w_I^2)$ and $\mathrm{Cov}(\gamma_I, \gamma_I \omega_I^2)$ are positive. However, in the two expressions for asymptotic power on page 12, this is not the only difference. For example, there is also $\mathrm{Var}(\gamma_I)$ versus $\mathrm{Var}(\gamma_I^{-1})$. Therefore, in the end, which one is more powerful is not completely obvious to me. Even in the simple case of $w_i = 0$ for $i = 1, ..., n-q$, I am not certain which test has better asymptotic power. I think a formal result comparing the power of the two test would be nice even under restricted assumptions.

Answer: We follow the reviewer's suggestion and add a formal result (Proposition 3) which gives the local asymptotic power of the proposed test and the test of Goeman et al. (2006). The local power functions are much simpler than the global power functions and make the comparison of the two tests clear. Let $\mathrm{Cor}(\cdot, \cdot)$ denote the Pearson correlation coefficient of two random variables. We show that the asymptotic local power of the proposed test is

$$\Phi\left(\Phi^{-1}(\alpha) + \phi\sqrt{\frac{n-q}{2}}\,\mathrm{Var}(\gamma_I w_I^2)\mathrm{Cor}(-\gamma_I^{-1}, \gamma_I w_I^2)\right),$$

and the asymptotic local power of the test of Goeman et al. (2006) is

$$\Phi\left(\Phi^{-1}(\alpha) + \phi\sqrt{\frac{n-q}{2}}\,\mathrm{Var}(\gamma_I w_I^2)\mathrm{Cor}(\gamma_I, \gamma_I w_I^2)\right).$$

Then it can be seen that the difference between the two tests lies in the behavior of the quantities $\mathrm{Cor}(-\gamma_I^{-1}, \gamma_I w_I^2)$ and $\mathrm{Cor}(\gamma_I, \gamma_I \omega_I^2)$. And the signs of these two quantities determine if the tests have nontrivial power. In the original paper, we have shown that the positive definite subspace for $\mathrm{Cor}(-\gamma_I^{-1}, \gamma_I w_I^2)$ is larger than that of $\mathrm{Cor}(\gamma_I, \gamma_I \omega_I^2)$. This shows that the proposed test can detect the signals from more directions than the test of Goeman et al. (2006).

2 Response to reviewer 2

1. The title should be changed. Nonexistence of unbiased test is not the main focus of this article, it is just the motivation of the paper. I suggest to change the title

to "A Bayesian test for high dimensional linear regression models with fixed design matrix".

Answer: We agree that the original title is not good enough. We partly follow the suggestion of the reviewer and change the title to "A Bayesian-motivated test for high-dimensional linear regression models with fixed design matrix". We preserve the word "motivated" since Bayesian method is just the motivation of our test and our test is actually a frequentist test.

2. The Introduction part should be re-written. The main focus of this article is two-aspects. One is the Bayesian test, and the other is the fixed design matrix. The introduction part should clearly state about your contribution through the two aspects compared with other competitors.

Answer: We follow the reviewer's suggestion and largely revised the introduction part. Hopefully, the revised paper can make clear the problem setting, our motivation and our contributions. We explicitly list our contributions in the introduction part, as follows.

> This paper has made three main contributions for testing hypotheses (2) in large p, small q setting with fixed design matrix. First, we prove that no test can detect all large deviations from the null hypothesis. This phenomenon is somewhat surprising as such test does exist in the random design setting. Second, we propose a novel Bayesian-motivated test, which can be regarded as an extention of the likelihood ratio test in high-dimensional setting. The good power behavior of the proposed test is verified both theoretically and numerically. Third, we propose to use Lindeberg's replacement trick to approximate the distribution of the proposed test statistics. The approximation is valid under weak conditions.

3. The transition through Section 2 to Section 3 is not clear. When you considering tests with good average power. Why Bayesian methods are natural choices? Some discussions are needed. In addition, the non-existence of nontrivial unbiased test is a trivial result for high dimensional setting.

Answer: We follow the reviewer's suggestion. In the revised paper, we improve the transition through Section 2 to Section 3 and add some discussions.

In frequentist point of view, Bayesian methods aim at minimize the average risk; See, e.g., Lehmann and Romano (2005), Section 1.6. Hence Bayesian methods are natural choices to produce tests with good average power. We make this clear in the revised paper.

The reviewer thinks the non-existence of nontrivial unbiased test is a trivial result for high-dimensional setting. In the revised paper, we also point out that our Theorem 1 implies the non-existence of minimax test. We do not think the non-existence of nontrivial unbiased test is a trivial result for high-dimensional setting. In fact, nontrivial unbiased test and minimax test do exist for the random design setting; see, e.g., Ingster et al. (2010). We point this out in the revised paper. Furthermore, nontrivial unbiased test and minimax test do exist in some other high-dimensional testing problem, including mean vector testing (Cai et al. (2013)), covariance matrix

testing (Cai and Ma (2013)). Hence we think our result in Section 2 is worth to be known.

We add the following transition paragraph at the end of Section 2:

Note that unbiasedness and minimaxity are two key considerations of frequentist hypotheses testing; see, e.g., Lehmann and Romano (2005). In our problem, however, these considerations can not be applied and there is no test with guaranteed power. Nevertheless, it is still feasible to propose a test with good average power behavior. Note that from the viewpoint of the decision theory, Bayesian methods aim at minimizing the average risk; see, e.g., Lehmann and Romano (2005), Section 1.6. This motivates us to propose a frequentist test with the aid of Bayesian methods.

4. Your proposed test is quite similar with that of the likelihood ratio test, some discussions for the differences are needed. In addition, since $\mathbf{X}_b^*$ is of dimension p, when both n and p are large, the matrix $\mathbf{X}_b^* \mathbf{X}_b^{*\top}$ is not invertible. Does your method applicable for $n > p$?

Answer: In Assumption 1, we assume $\mathrm{Rank}([\mathbf{X}_a; \mathbf{X}_b]) = n$, which implies $p + q \geq n$. Hence in this paper, we only consider the high-dimensional setting $p + q \geq n$. The reviewer's observation is correct, our test is not applicable for $n > p + q$. On the other hand, the likelihood ratio test can only be applied for $n > p + q$. Hence the proposed and the likelihood ratio test have the different applicable scope.

Indeed, the proposed test has an intrinsic connection with the likelihood ratio test. We discuss this connection at the end of Section 3.1:

In fact, if we apply the proposed methodology in the low-dimensional setting, that is, letting $\kappa \to 0$ in (4), then the resulting test statistic is exactly the likelihood ratio test. In this view, the proposed test is an extension of the likelihood ratio test to the high-dimensional setting.

We think this is an advantage of out test.

In comparison, the test of Goeman et al. (2006) is applicable for both low-dimensional setting and high-dimensional setting. However, in the low-dimensional setting, their test statistic is different from the likelihood ratio test. This property is undesirable.

5. The differences between your test and that of Goeman et al. (2006) and Lan et al. (2014) need to be discussed carefully.

Answer: The test statistic of Lan et al. (2014) is basically the same as the test statistic of Goeman et al. (2006). The difference of these two work mainly lies in the way of determine the critical value. Hence in our paper, we only compare the proposed test with the test of Goeman et al. (2006).

Following the reviewer's suggestion, we discuss the difference between the proposed test and the test of Goeman et al. (2006) more carefully in the revised paper. First, the two test methods use different strategies to set prior distribution of $\boldsymbol{\beta}_b$. This is discussed in Section 3.1. In fact, Goeman et al. (2006) let the prior magnitude of $\boldsymbol{\beta}_b$ tend to 0 to obtain a test with good local power, while

we let the prior magnitude of $\boldsymbol{\beta}_b$ tend to ∞ to obtain a test with good global power. Second, the two tests have different power behavior. In Section 3.3, we derive the asymptotic power functions of the two tests and show that the proposed test can detect the signals from more directions than the test of Goeman et al. (2006).

6. Throughout the article, I do not see any results related to fix design matrix, why the fix design is essential for your analysis? Is it useful for your theoretical results?

Answer: Fixed design matrix plays an important role in our Theorem 1 and the conclusion of our Theorem 1 will not hold for the random design matrix. In fact, as we have pointed out in the answer of Question 3, nontrivial unbiased test and minimax test do exist for the random design setting. Hence frequentist considerations can well be applied for the random design setting (most frequentist test for high-dimensional linear model adopt the random design matrix). In this view, our Bayesian-motivated test methodology is designed for fixed design matrix.

The theoretical analyses of the proposed test (including the determination of the critical value and the analysis of the power) are also carried out within the fixed design framework. Nevertheless, these results are still valid for random design matrix from a conditional inference perspective.

3 List of major changes

- We change the title to "A Bayesian-motivated test for high-dimensional linear regression models with fixed design matrix".

- We modify and re-written many paragraphs to improve the presentation of the paper.

- We add a proposition (Proposition 3) which gives the asymptotic local power functions of the proposed test as well as the test of Goeman et al. (2006).

- We add the simulation results of the test of Zhang and Cheng (2017).

References

Bai, Z., Pan, G., and Yin, Y. (2018). A central limit theorem for sums of functions of residuals in a high-dimensional regression model with an application to variance homoscedasticity test. *TEST*, 27(4):896–920.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372.

Cai, T. T. and Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819.

Goeman, J. J., van de Geer, S. A., and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493.

Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4(0):1476–1526.

Lan, W., Wang, H., and Tsai, C. L. (2014). Testing covariates in high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 66(2):279–301.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* Springer, New York, 3rd edition.

Wang, R. and Xu, X. (2020+). Least favorable direction test for multivariate analysis of variance in high dimension. *Statistica Sinica*.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768.